tremendously.

Is anything gained by converting the test from the original open-ended, or recall, version to an objective test format such as multiple choice? The objective test will be easier to score and with care the scoring can be accomplished with zero error, but subjectivity in the writing and interpreting of the key for the recall version has been pushed back into the selection of the misleads. By selecting misleads that capitalize upon fine nuances of meaning, the test can be made very difficult, and the subjects may appear to have restricted vocabularies. On the other hand, by selecting misleads that require only the grossest of discriminations, the test can be made quite easy.

It would be easily possible by any method of test construction to obtain three vocabulary tests with quite different distribution characteristics in the same population of subjects. For example, one could obtain means and standard deviations approximating those in the following table for each of three randomly selected 100-item tests with just a little trial and error.

	Test A	Test B	Test C
Mean	25	50	75
Standard Deviation	20	25	20

It is also reasonable to assume that there will be no gaps in the distributions of scores and that most of the possible range of scores will be represented on each of the tests. Furthermore, in large samples of subjects, say 1000 or so the distributions, whatever their shape will appear quite regular. One does not need third and fourth moments of the three distributions, furthermore, to draw inferences about the shapes of the three distributions. Test A and C are skewed, though in opposite directions, while Test B, though probably symmetrical in distribution, is more platy-kurtic than the normal distribution. These inferences all follow from the relationships of the standard deviations to the means.

Characteristics of Ordinal Scales. The linear intercorrelations of the three tests may be quite high--even a moderately well constructed test of 100 items for a high school population should be quite reliable--but these correlations can not be as high as the test reliabilities would allow. Some high scoring people on Test A, each with different total scores, will have identical scores on Test B and even more of them will have identical scores on Test C. A similar finding for low scoring people on Test C will be noted in comparing their scores on Test B and A. Such cases arise from the differential skews of the three distributions and regressions are necessarily curvilinear. In general the number of units by which any two scores differ in one distribution will be different than the number of units by which scores comparable in rank order differ in either of the other distributions.

It is also clear that ratios of scores computed for any one of the three tests are uninterpretable. The number 50 is twice 25, but getting 50 items right vs. getting 25 items right does not have the same meaning for each of the tests. Zero, furthermore, is quite a common score on Test A, is much less frequent on Test B, and is very rare if it occurs at all on Test C. A score of zero on any test would not indicate that the subject had a vocabulary of zero length. The accidents of sampling from the population of words are involved, but more importantly the arbitrary decisions of the test constructer serve to make a score of zero meaningless with respect to absolute size of vocabulary. The only information about a score of zero furnished by the test is that it is smaller than one.

The preceding characteristics of the three test scores define ordinal scales of measurement. The information furnished is basically rank-order information even though the numbers used have the appearance of an interval scale. The rank orders of subjects inferred from the numbers are not identical for the three tests, even allowing for measurement error, because the differences in skew will produce tied ranks on one test that are not tied on another. If we utilize the rank-order

information, however, and convert the obtained, raw scores to standard scores by means of a monotonic nonlinear transformation, i. e., by working through the percentile ranks, we can increase the linear correlations among the three tests as compared to linear raw score correlations.

Other examples will be presented to develop the argument in more detail and in more generality, but for the present the vocabulary test example can be taken on faith to represent the general case. Distributions of test scores are arbitrary. Tests furnish rank-order information only. Since a normal distribution has a number of desirable statistical properties, it is recommended that raw scores on tests be converted to normal distributions by means of the nonlinear transformation involving percentile ranks in a random sample of some defined population. When this has been done, the scale of measurement is said to have been normalized, but it is still ordinal. It has not been converted to an equal interval scale simply by means of the transformation. The choice of the normal curve conversion is a matter of convenience not of scientific necessity or conformity to natural law. If convenience dictates a different type of distribution, e. g., quartiles, deciles, or centiles for the converted scale, a different type should be used.

The fact that measurement with a test is ordinal is only the most obvious characteristic of this form of measurement. It is far from being the most important. It will become clear later that ordinal measurement has little effect on reliability or validity. Most of the inferences from test scores that are barred by the lack of interval or ratio scales are relatively unimportant and substitutes are generally available.

Possible Functions of Multiple Items. In the example of the preceding section tests of 100-items were assumed. One hundred different words were selected at random from an unabridged dictionary and subjects were asked to define these or to select the correct alternative from a list of misleads. This suggests a property

of the test that is of the utmost importance.

Tests are typically compsed of multiple items or "hurdles" with the subject behaving in some fashion with reference to each item. In measuring ability the performance required is either right or wrong, but in measuring personality or interest answers are frequently yes or no, like or dislike, etc. The total score on the test is also typically a linear combination of the scores on the items and, in many tests, weights for each item are either zero or one. It is not an essen tial feature of the test that the scoring be dichotomous, although dichotomous scoring is found very frequently. It is also not essential that the combination of items be linear, but nonlinear combinations can be dismissed from this discussion on grounds that such combinations are used only infrequently for research purposes and rarely if at all in standardized tests. The theory to be developed will assume a linear combination of dichotomous items for the sake of convenience and for wide-spread applicability, but the theory is directly applicable to all other types of item scoring with only minor modifications. Major modifications would be necessary, however, to adapt it to nonlinear combinations.

For the person steeped in traditional measurement theory the first hypothesis concerning multiple items is that they are required for purposes of reliability. All measurement is subject to some degree of measurement error. A scientist or engineer frequently makes multiple independent readings of his measures and uses the mean of these as his best estimate of the "true" value. Does the test differ in any respect from the need to take multiple measurements to reduce error?

There is, indeed, a difference in the practices of the engineer and of the psychologist. The person using a psychological test, when he wishes to increase his precision of measurement, repeats the whole test or uses parallel forms of the original test to obtain his sample of measures for which the mean is the best estimate of the subject's "true" score. The total test score is considered the

measure, not the score on an item. It is true that increasing the number of items in the test generally has the effect of increasing the reliability of the total score, but increasing reliability is not the primary function served by use of multiple items.

A second hypothesis concerning the function of multiple items is that they are required to furnish a scale more nearly approximating a continuous scale of measurement than does the dichotomous item. There are occasions, however, when it is not merely desirable but necessary to add together multiple items each of which is measured on a continuous, equal interval scale, i. e., total score composed of multiple items may be required even though the items are not dichotomous. Again, multiple dichotomous items do furnish a scale approximating a continuous measure, but this is secondary to their primary function.

The Important Function Served by Multiple Items. The principal function served by multiple items is best seen as a contrast between test theory per se and traditional measurement theory. In the ordinary measurement of height it is reasonable to assume that each measurement operation for each person measured includes a true score component and a random error component. This is the starting point for classical measurement theory. The variance of obtained scores includes the variance of true scores and the variance of error. Correlations involving the obtained scores are a function of the covariances with true scores and the variances of obtained scores. From this basis such statistical concepts as the standard error of measurement and correction of correlations for attenuation by measurement error are readily developed.

Classical measurement theory is not, however, readily applicable to the test. Efforts to use classical theory over the years, furthermore, may have hurt test development as much as it has helped. The major departure from classical theory arises from the necessity to start with a definition of item score that differs

significantly from classical theory. An item score is composed, as Loevinger has discussed (1954) most fully, of three distinct parts: score on the trait or disposition (d) in which the test constructer is interested, systematic nonerror noise or bias elements or factors (b), and random error (e). The important effect of using multiple test items is to minimize the effects of the numerous nonrandom factors that are subsumed under the label of noise.

To return to the vocabulary example, a high school student may have encountered a word in his recent reading for which he obtained a definition and when this word was encountered on the test he answered it correctly. The word may be difficult in general and the student's general vocabulary level low, but he obtained an extra point in his total score for a nonrandom reason independent of his general level of vocabulary competence. There are many such examples. Some words are encountered more frequently in science than in the humanities, in pulp magazines than in school books, or in certain neighborhoods or social levels than in others, and so on. By taking a large sample of words such effects, although still present, can be balanced off against each other, and the more general disposition to know the meanings of words will be measured with greater validity.

In this connection it is instructive to look at item intercorrelations for some standard ability tests. In tests that are quite homogeneous both with respect to difficulty level of the items and the subject matter of the test, item intercorrelations with a mean as high as .20 are not common in the full range of talent and occur quite rarely in special groups who are restricted in range of talent. While it is not easy to assess the random error component of variance in an item--memory for item content makes suspect a repeated measures design and parallel items are not easy to construct--it is probable that nonerror or bias factors are a major contributor to item variance for most of the items that appear in psychological tests.

The vocabulary example suggests an alternative designation of nonrandom noise or bias in test items as item sampling error. This is indeed one source of bias, but the equivalence breaks down in two important ways. In many, many cases there is no defined population of items from which to sample, though the use of item selection error could get around this difficulty. A more important difficulty, however, is that certain bias factors are intrinsic to psychological items. Every test item has a particular item format, a time limit or work limit, a set of directions. In addition, each examinee has a different background of knowledge, skills, sets, and other experiences. All contribute variance to a test item. Reasoning is necessarily measured with verbal symbols, numerical symbols, or figural materials. Words that appear in a vocabulary test occur with differential frequency in different kinds of reading material. The use of noise or bias suggests unwanted or even uncontrollable, which is desirable, and the use of systematic indicates that the behavior measured is lawful, which is also desirable.

Just as a weed is any plant growing where it is not wanted, systematic noise or bias includes any factor or element appearing where it is not wanted. One man's noise, for one purpose, becomes another man's primary mental ability, for another purpose. But unlike weeds, a great deal of systematic bias is impossible to eradicate.

The Correlation Between Test and Criterion. A little algebra may be helpful at this point. Let there be n items in the test and let d, b, and e represent the disposition the test constructor desires to measure, the bias factors, and random error, respectively. The correlation of the test of the disposition with a criterion measure of the disposition is given by the following:

$$r_{xy} = r_{(x_i + \ldots + x_n)y} \text{ and each } x_i = d_i + b_i + e_i$$

$$r_{xy} = \frac{\Sigma\, Cd_{iy} + \Sigma Cb_{iy} + \Sigma Ce_{iy}}{s_y \sqrt{\Sigma s^2_{d_i} + \Sigma s^2_{b_i} + \Sigma s^2_{e_i} + \Sigma\Sigma\, C_{d_i d_j} + \Sigma\Sigma\, C_{b_i b_j} + \Sigma\Sigma C_{e_i e_j} + \Sigma\Sigma C_{d_i b_i} + \Sigma\Sigma C_{d_i e_i} + \Sigma\Sigma C_{b_i e_i}}}$$

If error is truly random, then all covariance terms involving e will drop out. Psychometricians have typically been willing to make this assumption. Moreover, if the b terms were specific to each item and independent of the disposition score on the item, they would have the functional characteristics of error and covariance terms involving b would also drop out. While it is not difficult conceptually to assume orthogonality of disposition and bias factors, the assumption of specificity of bias to each item is almost always false.

It is also reasonable to assume that the noise factors are unrelated to the criterion measure. (This can be considered true by definition.) With these considerations in mind, formula 1 can be rewritten as follows:

$$r_{xy} = \frac{\Sigma C_{d_{iy}}}{S_y \sqrt{\Sigma S^2_{d_i} + \Sigma S^2_{b_i} + \Sigma S^2_{e_i} + \Sigma\Sigma\ C_{d_i d_j} + \Sigma\Sigma\ C_{b_i b_j}}}$$

In the best of cases bias factors are minor sources of variance of total score on the test (denominator) and make zero contributions to covariance with the outside variable (numerator) currently of interest. In the worst of cases the nonrandom bias variance of the test is entirely noise from the point of view of the aims of the test constructer, and the only nonzero terms with criteria involve sources of variance other than the one the test is supposed to measure.

By basing the total score on many items, it is possible to build up the validity of the test for a particular disposition even though any one item has only a small component of that disposition. The secret is to spread item selection over as many bias factors as possible so that any one bias factor runs through a minimum number of items. The goal, though frequently unattainable, is to make the bias factors specific to items. Even when it is impossible to keep the bias covariance terms near zero in the denominator, the scattering of this

variance among many bias factors will avoid the situation in which the total score is a better measure of some other disposition than the one intended. Many items, therefore, are a necessary though not a sufficient condition for building up the variance of a particular disposition in the total score on the test. A basic misconception concerning the original choice of items will result in the test constructer measuring something, with greater and greater precision as he continues to add items, that he does not wish to measure.

INTERRELATIONSHIPS OF HOMOGENEITY, RELIABILITY, AND VALIDITY

The concept of homogeneity of a test does not appear in classical measurement theory. Homogeneity with respect to content is an issue only in those situations in which multiple items are used. The statement made earlier that multiple items did not serve the same primary purpose as multiple measures in physics or engineering, but did have a secondary effect on reliability is important in this connection. Many psychologists are confused on this issue. Homogeneity interacts with both reliability and validity, but must not be confused with either.

Homogeneity and Reliability. Kuder-Richardson homogeneity coefficients are frequently called reliability coefficients. Under certain restricted circumstances, it is true, one can obtain a reliability estimate from a measure of the homogeneity of the test, but it is essential that the investigator keep the distinction, and the conditions, clear in his own mind and in his writing.

The Kuder-Richardson formula best used to estimate reliability is the following:

$$r = \frac{n}{n-1} \frac{S_x^2 - \Sigma P_i q_i}{S_x^2}$$

Only the number of items, the difficulty levels of the items, and the variance of the total score on the test are used. (This is algebraicly equivalent to the approach to homogeneity of a set of measures by means of the analysis of variance which Hoyt (1940) suggested.) The variance of the total score is, in turn, a function of the number of items, the item variances, and the item covariances. These parameters do not have a one-to-one relationship to reliability defined as the correlation between repeated measures or between repeated parallel measures.

The difference, and the relationship, between homogeneity and reliability can best be shown by writing out the formula for the repeated measures correlation as a function of the relationships involving the items. (The prime refers to the repeated or parallel item or test total score.)

$$r_{xx'} = r_{(x_1 + x_2 + \ldots + x_n)\,(x'_1 + x'_2 + \ldots + x'_n)}$$

$$r_{xx'} = \frac{\Sigma C_{x_i x'_i} + \Sigma\Sigma\, C_{x_i x'_j}}{\sqrt{\Sigma S^2_{x_i} + \Sigma\Sigma\, C_{x_i x_j}}\ \sqrt{\Sigma S^2_{x'_i} + \Sigma\Sigma\, C_{x'_i x'_j}}}$$

If item intercorrelations are zero, the right hand term in the numerator disappears as do the right hand terms under each of the radicals. The reliability coefficient is then completely a function of the item reliabilities and can vary from zero to one. Furthermore, one can conceive of a test in which these conditions would be rather closely approached. A scored biographical data blank, for example, could contain items that were essentially uncorrelated with each other, but the reliability of answering an individual item would be very high.

As the intercorrelations of items within and between tests approaches the correlations between the paired items, the Kuder-Richardson homogeneity coefficient approaches the reliability coefficient of the test. For the two to be equal in conception the item difficulties would all have to be the same. Otherwise covariances are necessarily somewhat attenuated. In practice, this latter condition can be ignored since the formula does take out the variance due to the main effect of difficulty level, and variations of difficulty level within the normal range have only a slight, biasing effect on the interaction between persons and items which is the essential term determining the homogeneity of the test.

The interaction between reliability and homogeneity is more clearly seen if Formula 5 is rewritten to make explicit the assumption that the retest or parallel measure is identical with the first; i. e., test variances are equal and intercorrelations within tests are equal to intercorrelations between tests.

$$r_{xx'} = \frac{\Sigma C_{x_i x'_i} + \Sigma\Sigma\, C_{x_i x_j}}{\Sigma S^2_{x_i} + \Sigma\Sigma C_{x_i x_j}}$$

The right hand quantities in the numerator and denominator are identical. The left hand quantities represent the ratio between paired item covariances and variances. With many items in a test the right hand quantities will generally be much larger than the left hand ones; the homogeneity of the items, in other words, typically makes a larger contribution to reliability than the reliability of the paired items. The test constructer by narrowing the focus of his test, i. e., by redefining the disposition in which he is interested to make it coincide with an important source of nonerror noise, can step up test reliability very easily. To suggest that this may be undesirable may seem strange to those imbued with classical measurement theory. Why should not the ratio of true score variance to total variance be maximized? The answer is, of course, that an increase in reliability is not worth the price if the disposition which the psychologist seeks to measure is redefined in the process of test construction to make it less useful psychologically.

As a matter of fact, the positive steps in test construction that follow from the concept of disposition, nonrandom noise, and error contributions to item variance make it difficult to achieve high reliability with a limited number of items. The variance of noise or bias factors must be spread around as widely as possible. The more successful the test constructer is in his efforts, the lower will be the item covariances. He can compensate for this effect only by increasing the number of items in the test.

Homogeneity and Validity. No one administers a test, however, simply to obtain reliable information of some sort about a person. Tests are administered in order to make inferences about behavior: inferences about jobs, school, military assignments in applied work, or inferences about functional relationships involving a particular disposition in more basic research. Validity is a shorthand term used to cover the inferences that can be drawn from a score. Validity

coefficients stated in terms of item characteristics were presented in Formulas 1 and 2, but a simpler one will now be more convenient. This one is stated in terms of the items, without regard to their components, and their relationship with any outside variable, y.

$$r_{xy} = \frac{\Sigma C_{x_i y}}{S_y \sqrt{\Sigma S^2_{x_i} + \Sigma\Sigma C_{x_i x_j}}}$$

Formula 7 shows that, item validities being equal, there is a premium placed on low homogeneity. Item covariances occur only in the denominator. High reliability which comes about through an interaction with homogeneity is indeed a misleading goal. Only in case certain subsets of items in a heterogeneous test have zero correlations with the criterion does it pay to increase the homogeneity of the test and obtain the concomitant increase in reliability. When all items are related to the outside variable, by keeping item intercorrelations low, the variance of the test score will be kept low and reliability will be kept low, but the size of the validity coefficient increased. Such reasoning is completely compatible with expectations based upon multiple regression theory, but it does require qualification of the classical theory concerning the relationship of reliability to validity.

Reliability and Validity. Classical theory states the relationship of reliability to validity in the correction for attenuation.

$$\hat{r}_{xy} = \frac{r_{xy}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}}$$

This formula is applicable to test theory only in cases where reliability is changed by the addition or, with appropriate changes in the formula, subtraction of items of the same type as the originals. Whenever items are discarded

selectively with others retained, an increase in reliability may accompany a decrease in the validity of the test. Increasing the reliability of a test by doubling its length with exactly comparable items will increase the test's validity. Increasing the reliability of a test by item selection procedures will not have a predictable effect on the test's validity.

The same assumption, i. e., adding exactly comparable items, must be made in estimating the reliability of a test of a different length than the original, but the importance of the assumption in this case is better known. It may be instructive, however, to apply it to the hypothetical situation described earlier in which item intercorrelations in the original test are zero.

When item intercorrelations are zero, test score reliability is more or less the mean of the item reliabilities. When the length of the test is doubled by the addition of exactly comparable items, item covariances are no longer zero. The assumption of comparability means that each item in the original test now has itself or a parallel version of itself in the test of increased length. The new test no longer has zero homogeneity.

Minimum Requirement for Homogeneity. Even though completely uncorrelated items would be best to maximize the correlation between a test and an outside variable, such a set would not be considered to measure a psychological disposition. It is here that the concept of the homogeneity of the test is required. Various indices of a disposition of psychological interest just ought to have something in common. If the disposition of glass to shatter or beams to snap under the stress were measured in a fashion analogous to the psychological test, the various items would be correlated just as the items that measure height or weight in the other physical analogues to the test are correlated.

If the reality, or even the necessity, for some degree of homogeneity is accepted, it does not necessarily follow that homogeneity should be as high as

possible within the limitations of obtaining a measuring device furnishing a nearly continuous score. (If a test of height were given very reliably, maximum homogeneity would result in a U-shaped distribution. Some degree of spacing of item difficulties, with the resultant decrease in item covariances, is necessary in a test for height to discriminate among the examinees.) Some degree of homogeneity is expected, but the degree is optional. The degree depends both upon the psychological facts, i. e., the extent to which behavior is dispositional as opposed to situational, and upon the breadth of the disposition that the psychologist wishes to measure. He may be interested in measuring intelligence, or perhaps something even broader than intelligence, or at the other pole in measuring the fluency with which four letter words beginning with s can be evoked.

Let the rule be that any set of positively intercorrelated items can be added together. Such composites produce a psychologically meaningful total score, particularly if the items are intercorrelated at about the same level. It does not matter whether this level is low or high. If one is trying to measure some disposition that is very broad, and in consequence each item may contain only a very small portion of the variance of the disposition, it will be necessary to plan on using many items widely scattered in order to dissipate the many possible sources of nonrandom noise. That other factors will contribute to the total score is of no consequence. As a matter of fact, the larger the number of these the better since this will tend to keep the contribution of each small. The restrictions that item intercorrelations be at about the same level is necessary in order to avoid giving undue weight to a particular bias factor. It can, however, be relaxed if this is carefully done. The bias factors must themselves be evenly distributed.

The Goal of High Homogeneity. In contrast to the preceding rule, those who set high homogeneity of the test as their goal, imply that only those items can be

added together that have the very highest level of intercorrelations. If any given test can be broken down into subsets of items whose intercorrelations are a little higher than the cross correlations between subsets, the original test is "impure" or heterogeneous and new tests should be constructed as defined by the subclusters of items.

The reader who believes the rule that any set of positively intercorrelated items can be added together is ambiguous--after all there are many possible levels of item intercorrelations and thus many possible tests--should ponder the ambiguity in the high homogeneity rule. How high is the highest possible level of intercorrelations? When does the correlation between two different items become sufficiently high that it should be considered the correlation between parallel forms of the same item? If this rule is pushed to the extreme, does it not mean that the ultimate in homogeneity is reached when one reaches a small set of items that are essentially parallel forms of each other?

In discussing Binet's interest in multiple intellectual functions and the development of the Binet scales of intelligence, Guilford (1967) concluded that Binet's decision to use a single score for the totality of his items (mental age) was completely incongruous. In the light of the present discussion Guilford's conclusion is simply incorrect. One can accept multiple factors both of the Thurstone sort and of the Guilford sort, which appear to be narrower than the Thurstone primary mental abilities, along with a general factor without any logical or psychological difficulty. (See Humphreys, 1962, for a fuller discussion of this issue). Ability items are positively intercorrelated to varying degrees. High intercorrelations determine narrow factors; moderate intercorrelations determine somewhat broader factors; low intercorrelations determine a general factor. It is completely reasonable for an investigator to measure with a single test the factor or complex of factors that produce the lowest positive correlations

among broadly distributed items.

It is important to realize that the argument here is empirical, not logical. Good characteristics as defined socially of the human being tend to be positively correlated. The most disparate abilities, with abilities used in the most general sense, such as correlations between clerical and mechanical abilities, or between information about farming and about social sciences, are positively correlated. Psychological dispositions of the abilities sort are also positively correlated with physical measures such as height and weight. Terman's gifted children were healthier, wore fewer glasses, etc., on the average, than other children.

This tendency for "all good things to go together" is much more marked when one samples from men or women in general in a given cultural group than when samples are drawn from more restricted ranges of talent. Even within samples of college students enrolled in the most highly selective institutions correlations still tend to be positive though occasionally negative values occur which can generally be explained in terms of sampling errors. Negative correlations in a restricted population do occur, but there is frequently a sampling explanation. Highly selective universities can not play big-time football without having separate standards for athletes and nonathletes. Correlations between athletic abilities and intellectual abilities will be negative in such mixed groups.

Arguments Pro and Con. One argument advanced against the broad test is that "purer" tests are better than more complex tests on grounds basically of scientific esthetics. Here there is a difference in point of view as to what constitutes a pure test. Guilford's factor pure tests are seen by the present writer as inextricably complex. Tests of high homogeneity that measure one of the "aptitudes" in Guilford's structure of intelligence reflect simultaneously variance introduced by all of his three dimensions. Such tests are like the physical analogue in which weight was measured by having each subject lie down in a uniform manner at the end

of the lever: scores were highly homogeneous, but reflected both height and weight in an unknown combination.

A second argument against broad tests, and an almost convincing one, is that all of the information in the most complex test is basically available in a large number of highly homogeneous tests of the Guilford sort. Potentially also, information is lost by moving from many tests to a smaller number of broader tests. There are two different counter arguments to this point both of which are matters of feasibility. It would be very difficult to motivate examinees to sit through and work well for the amount of time necessary to administer 120 tests each sufficiently reliable to justify a separate score. It is also very difficult statistically to obtain stable weights for 120 measures for the various sorts of inferences in which psychologists are interested. It is not an exaggeration to estimate that the number of cases required would run into the tens of thousands for each outside variable considered. This estimate has a statistical basis in the formula for the standard error of a beta weight and an empirical basis in the ubiquity of positive intercorrelations among items and tests.

It is even possible given optimum Ns for weighting purposes, little if any information would be lost by the use of broad tests carefully constructed (Humphreys, 1962). Tests of the analogue to the main effects in an analysis of variance for each of Guilford's dimensions might well furnish the same information as the 120 tests representing the Cartesian product of his dimensions. The 120 different combinations of those dimensions become a source for selection of items and a guide for distributing noise factors as widely as possible in this conception, but not a mandate to construct 120 tests for assessment purposes.

ILLUSTRATIONS OF TEST CHARACTERISTICS BY MEANS OF PHYSICAL ANALOGUES

It is possible to develop physical analogues to the test that help to clarify the principles that have been presented. These principles will also be developed more fully as the various physical analogues are discussed.

Behavioral Test of Height. A carpenter is asked to make a series of standards in the form of an inverted L with the only specifications being that the uprights will all differ from each other and that they will cover the range in height of adult men. It is not essential that the horizontal bar be at right angles with the upright, and the essential specifications are checked perceptually only. Each standard is given a separate designation, perhaps a number. A sample of men is drawn from a population; each man is confronted with each of the standards in turn in a uniform manner; and each man is given a score representing merely the number of times the tip of the horizontal bar touched his head when an attempt is made to pass it over his head with the upright being placed vertically on the ground.

If a very large number of standards is constructed initially, it should be possible to select from the larger group a smaller set having any specified distribution of item difficulties. (This statistic is readily computed: the number of heads hit by a standard divided by the total number of men in the sample measured provides a statistic varying from zero to 1.00 with high values representing "easy" items.) If 9 standards are selected having difficulty levels ranging from .10 to .90 by steps of .10, the distribution of total scores for height will be rectangular in shape; i. e., symmetrical but highly platykurtic.

It is not difficult to see how the shape of the distribution of total scores can be inferred from the distribution of item difficulties. The easiest item has a difficulty index of .9, i. e., 10% of the sample fails the easiest item. Their total score on the test will be zero. Another 10% fail the item having an index

of .8. While 20% fail this item, one-half of the failing group passed the easier item. Thus 10% will have a score of 1. In a similar fashion the 10% who passed the most difficult item passed all easier items. Since this is the 9th standard in order of difficulty, only 10% of the sample will have a score of 9.

With 10% of the sample at each score, the distribution is rectangular.

It is instructive to make a table in which the items are placed in a horizontal array in order of increasing difficulty and the subjects are placed in the vertical array in order of increasing size of total score. (In this example only, all subjects having the same score can be represented by a single tally.) The result, in Table 1, is a triangular matrix of tallies which defines what has become known as a perfect Guttman scale. No man fails an easier item after having passed a more difficult one. When the number of tallies in a column is counted and divided by the number of people, the difficulty level of the item is the result. When the number of tallies in a row is counted, the total score on the test is a result. (A percentage score on the test is sometimes obtained by dividing the total score by the number of items in the test, but the number of items and the zero point are quite arbitrary.)

It is of interest that the product-moment intercorrelations of the items in a perfect Guttman scale form what Guttman has called a simplex matrix (Guttman, 1955). The simplex matrix indicates the presence of a single underlying function or factor when the successive variables differ in difficulty level, complexity, growth, or level of learning (Humphreys, 1960). The intercorrelations of the present example are presented in Table 2 for purposes of illustration. It should also be noted in connection with this example that the presence of a single factor is inferred from the form of the correlational matrix and the known differences in difficulty level of the items. The simple, one factor explanation cannot be ob-

tained by the application of the usual factor analytic methods. If squared multiple correlations are alternated with unities in the diagonal,

the 9 items will define four principal component factors (Humphreys, 1960).

What would the test constructer do if he wished to obtain a normal distribution of test scores for his measure of height? He would go back to his population of standards and select those having an appropriate distribution of item difficulties. Difficulties ranging from .96 through .89, .77, .60, .40, .23, .11, to .04 would produce a distribution of total scores that would be approximately normal. The mean would be the same as the mean of the rectangular distribution, but the standard deviation would be smaller.

The test constructer by appropriate selection of item difficulties can produce a distribution having any shape he desires. Wide variations in both kurtosis and skewness are possible. A test for the selection of basketball players can be produced having a tail at the upper end of the score distribution. After all, the coach is not concerned about making discriminations among college freshmen who are in the lower quarter, or even half, of the distribution of height. U-shaped distributions are possible though hardly useful. For a general purpose test, however, the test constructer does not worry very much about the distribution of item difficulties needed to produce a raw score distribution having a particular shape. Instead he takes what the accidents of item selection produce and converts the raw scores, by means of a nonlinear, monotonic transformation into a distribution of converted scores. As indicated earlier, it is frequently convenient that the shape of the converted score distribution be normal.

A very important reason why the test constructer does not worry too much about the selection of item difficulties in most cases is that psychological items do not readily scale in the Guttman sense. One reason why this is true is the

unavoidable presence of measurement error in each item. While the presence of the Guttman scale made possible the characterization of the shape of the score distribution from knowledge of the distribution of item difficulties alone, it was also necessary to assume error free measurement in order to obtain that scale. The measurement situation had to be highly standardized. Instructions to subjects were given to control posture; instructions to the test administrator controlled the nature of the surface on which the subject stood and the placement of the standards relative to the subject.

The Introduction of Measurement Error. This principle can be illustrated by returning to the test of height and the population of standards originally postulated. The only change to be introduced is that uniform conditions of measurement will not be specified. Posture will no longer be controlled. Neither will the measurement surface nor the placement of the standard be controlled. All of these will be allowed to vary at random from subject to subject and from item to item within a subject.

When nine items are now selected having the same equally spaced distribution of item difficulties as before, the distribution of total scores will no longer be rectangular, but instead will be unimodal. When subjects and items are tabled as before, many persons will be found who have failed an easier item after having passed a more difficult one. Something like the item data in Table 3 will be the result though Table 3 is schematic only if it is taken to represent more than 10 subjects. The number of subjects who fail easier items after passing, will be functions of the amount of error that has been introduced by the failure to standardize the measurement situation.

Limitations of the Ordinal 4 Scale. The usual statistic expressing the reliability of measurement of a test is the correlation between repeated tests or between parallel forms of the test. If the conditions of careful, standardized

measurement did produce a perfect Guttman scale, the correlation between test and retest, or between two separate tests having identical item characteristics, would be unity. In a larger group of items occasional reversals would be found under even optimum conditions of measurement and the reliability coefficient would only approach unity. There is no essential reason, however, why the reliability of the test of height should not be every bit as high as the reliability of the usual measurement of height. It is all too easy, however, to be careless with any scale of measurement, and it is probably easier to introduce error into the test than into the use of a physical scale of measurement just because there are more occasions with multiple items for error to occur. Without standardization of the measurement situation, as in the second example, reliability coefficients will depart substantially from 1.00.

There is also no essential reason why the correlation between the test of height and the criterion measure of height should not approach unity. Lack of uniform conditions for the test, as well as carelessness in the measurement situation for the criterion, will attenuate the validity coefficient of the test, but there is nothing intrinsic to an ordinal scale that produces a reduction in validity. The only inferences barred are those involving equal units or equality of ratios and the absolute zero. For most purposes to establish converted scores in a meaningful population of subjects provides useful though not full substitutes for the standard deviation (requiring equal units) and the mean (requiring an absolute zero) of the ratio scale.

Probably the most important type of inference barred by an ordinal scale is the characterization of the form of the functional relationship between a psychological disposition measured by the test, as the dependent variable, and some independent variable. There is no point in worrying about power versus log functions, for example, if there is doubt concerning the equality of the units of measurement.

By making certain assumptions about the nature of human judgment it is frequently possible to get outside the limitations of the test as here defined and obtain equal units. Problems of scaling have been discussed thoroughly by Torgerson (1958). For present purposes it is sufficient to add that interval and ratio scales formed by such assumptions must be thoroughly and independently checked. Thus, the supposedly equal interval attitude scales of Thurstone and Chave (1929) do not have intervals that are equal independent of the attitudes of the judges who do the scaling. The lack of equality as a function of attitude of the judge is more marked for the equal appearing interval method of scaling than it is for the paired comparisons method, but it is not completely absent in the latter (see Edwards, 1957, for an extended discussion of these data). It is also true that a Likert type scale (Likert, 1932), which is clearly ordinal in the sense here described, is probably just as valid as a Thurstone scale (Edwards, 1957).

Determinants of Test Score Distributions. With respect to the shape of the distribution of test scores, two generalizations are possible at this stage of the development: (1) the variability of item difficulties is inversely related to the variability of the distribution of scores on the test, or is directly related to a change in the form of the distribution toward leptokurtosis. (2) The amount of error present in the testing situation is inversely related to the variability of the distribution of scores, or directly to a change in the form of the distribution toward leptokurtosis.

The second generalization above appears to be at variance with classical measurement theory. It is easy to prove in the classical theory or in the Loevinger variant of that theory that the variance of true scores or of disposition scores is always less than the variance of obtained scores. It is not always noted, however, that the preceding conclusions demand an interval scale of measure-

ment. For the test this means, when the same set of items is administered once carefully and once carelessly, that two different ordinal scales are the result. The set of items administered carefully will have the larger standard deviation, but the ratio of the variance of true scores to error will also be larger in that set. In contrast, when height is measured carelessly on the physical scale, the variance of the obtained measures is larger, and the ratio of true score variance to error variance is smaller, than when height is measured carefully on the same scale.

Another way of illustrating these principles is to write the standard deviation of the distribution of total scores on the test in terms of the item characteristics. The effects of dispersion of item difficulties and the introduction of error can be observed in the item statistics.

$$s_x^2 = \Sigma p_i q_n + \Sigma\Sigma C_{ij}$$

The largest contribution to total variance of the item variance terms is obtained when $p = q = .50$ for all items. The largest contribution to the item covariance terms, on the assumption that all covariances will be positive (all items are assumed to be measuring the same function), is obtained when $p = q$ for all items. Wide variation in item difficulties reduces the contributions of item variances and covariances. For most tests the covariance terms are more important than the variance terms just because there are so many more of them.

The effect of increasing the amount of random error in the measurement situation comes about by way of attenuation of the size of the covariance terms. Error decreases the size of correlations among obtained scores relative to true scores. The greater the amount of error, the smaller the size of the item covariance terms which reduces the size of the standard deviation of test scores. The variance of a test score distribution attenuated in size by the presence of error of measurement will contain a larger proportion of error variance relative to true score variance than will the larger variance of an error-free test of the

same number and distribution of item difficulties. The absolute size of the standard deviation is smaller, however, for the error-ridden test.

It is obvious also that the size of the standard deviation of test scores is a direct function of the number of items. The addition of items of any type, error-free or error-ridden, will increase the size of the overall standard deviation. The addition of even a single item to a test changes the scale of measurement.

The Introduction of Nonrandom Bias. In order to consider effects of nonrandom noise or bias it will be useful to construct another physical analogue to the test. A test constructer interested in measuring weight has a lever, a fulcrum, and a pile of big rocks. In a pilot study he finds a place for the fulcrum that will allow the typical rock to just about balance the typical male adult. (The use of mean rock and mean adult has been avoided to indicate that the pilot research does not have to be precise.) Again the rocks are each given an identification and each man in the sample is placed on the lever opposite each of the rocks in turn. The score is the total number of rocks raised in the air by the passive man. The test constructer has a partially incorrect theory about what he is trying to measure however, and carefully instructs his subjects to lie down on the lever with their bare feet precisely at the end and with their heads extending toward the fulcrum as much as necessary. (He has been thoroughly indoctrinated with the necessity for care in measurement.)

Proceeding as before, ten items are selected which produce a rectangular distribution and a perfect Guttman scale. Scores represent an unknown mixture of height and weight, but there are no data from the measurement operations alone that lead to this conclusion. As long as great care is taken in measurement, no man will fail an easier item after passing a more difficult one. The first generalization from this new example, therefore, is that systematic measurement of nonrandor

noise does not necessarily reveal its presence.

If the test constructer had instructed his subjects to lie down as described on a specified one-half of the items and to stand with their heels at the tip of the lever on the other half, scores would still represent an unknown mixture of height and weight, but there is a possibility that the presence of a second factor in the items could be detected. In this new example of measurement of systematic bias, there are two types of items which would be expected to show their differential similarities in their correlational patterns which would in turn determine two factors. Unfortunately, item correlations are affected by range of difficulties as well as by content (see Table 2 in this regard). Differences in item marginals may cloud the statistical differentiation between the two factors, but there is hope in this example in being able to show the presence of the two factors in the data. Even with careful measurement, when two factors are present in the items, a Guttman scale will not be obtained; some subjects will fail easier items after having passed more difficult ones.

Even if the two types of items are separated perfectly on the basis of information internal to the measurement operation, there is no statistical clue from the item data as to which is the better measure of weight. The proper identification of the function each cluster is measuring might be made intuitively from inspection of the items in the separate clusters, but external relationships represent a more dependable means of identification. Thus the factor analytic method is vulnerable on two counts: (1) the difficulties in factoring dichotomously scored items and (2) adequate identification of factors without pursuit of differential, external relationships.

If posture is allowed to vary from standing to sitting to lying down and if this variation occurs at random from subject to subject and from item to item, what was nonrandom noise becomes random noise. The effects of error have been

described earlier. There is, however, an in-between condition which is more serious. If posture varies from subject to subject but not from item to item, one subject's score may represent a relatively pure measure of weight while another's identical score may represent a mixture of height and weight. Again, as in the first example, there are no internal clues. A Guttman scale can be obtained under such conditions, for example, but information external to the measurement operation is necessary in order to identify those subjects whose measures are valid measures of weight. Without separation of subjects this type of nonrandom noise would depress validity coefficients. If subjects could be separated, however, validity would be very high in one sub-group, quite low in another.

Increasing Complexity of Bias Factors. Although a number of test construction principles have been demonstrated by means of physical analogues, up to this point in the development there has been no precise, complete analogue for the most important reason for the use of multiple items described earlier. Before introducing such an analogue it will be useful to return briefly to that argument.

Tests require the subject to behave on each of a number of items. An underlying trait or disposition to behave in certain ways is inferred from the test score. Yet there are myriads of possible causes for behavior. Any one bit of behavior may reflect the underlying disposition only in small degree. Knowledge of any one word does not indicate very much about a disposition to know many words. This phenomenon was described in terms of an analysis of the test score into disposition, nonerror noise, and random error components. Individual test items generally contain much more variance from nonerror noise and random error than they do from the hypothesized disposition.

A physical analogue that will illustrate this property can again be devised. A test constructer without a tape measure hopes to measure height. He also has no standards long enough to meet the criterion used in the first analogue example.

He can construct items that will measure the length of toes, fingers, lower arms, lower legs, and head; various measures of width and depth of arms, legs, trunk, and head are possible; various circumferences can also be measured. Such items could be dichotomous, but might even be measured by tape or calipers on the physical scale of measurement.

The first principle to note for this analogue is that, if dichotomous, the items would depart radically from a Guttman scale. Many, many failures on easy items after passing more difficult items would be evident. Long fingers do not typically accompany a broad chest. Also, for a given number of items, the standard deviation of the test scores would be lower than in previous examples for items of similar difficulty levels and with equal amounts of error. The distributions would be unimodal, even with minimum error in the measurement operations, and with little variability of item difficulties. Systematic noise or bias of this type, which is typical of psychological tests, reduces mean item intercorrelations substantially. The net effect on the test score distribution is similar to the effect of error. With sufficient item heterogeneity the test constructer does not have to distribute item difficulties in order to have a useful ordinal scale. Item difficulties clustered closely around .50 will produce a U-shaped distribution with highly homogeneous items carefully measured, but the same distribution of item difficulties with heterogeneous items carefully measured will produce a unimodal distribution of total scores.

Criteria for Item Selection. If there were an objective external criterion of height available, (a) it would be simple to obtain multiple regression weights for the set of items, (b) but under these circumstances it would be unnecessary to measure height with a set of items. The analogue to the test is to combine these items selectively in a linear fashion on the basis of internal data alone.

An equally weighted linear combination of all of the items listed would undoubtedly produce a score that would be rather highly correlated with height. This score would probably be more highly correlated with height than would any one of the components, but some sub-set of these items might produce an even higher correlation. Clustering or factoring might be possible, subject to the reservations expressed earlier about the effects of disparate item difficulties, but this approach is far from simple. Depending upon the density of sampling of bodily measures one might find both finger and toe length factors, a long bone factor, a body width factor or factors, circumference factors, etc. If only one factor is to be used, it is probable that the long bone one is most highly correlated with height, but the test constructer working without this knowledge would have difficulty justifying this selection. From item data alone his grounds could only be intuitive, following inspection of the items, and the items obviously refer to bone length and not to full stature. Furthermore, it is highly probable that the entire set of items carries more information about height than does the long bone subset, so the problem of the test constructer is to bring in all of the useful information and exclude the useless information from his test.

This situation can also be viewed in terms of the factor analytic methods with particular reference to the problem of factoring in several orders. With the use of very large numbers of items as contemplated in this test of height, there would probably be several first order factors defined by anatomical location and by the dimension, i. e., length, breadth, or depth, measured. These factors would contain relatively little information concerning full stature; instead they represent mainly systematic noise. Factors more closely related to full stature would be found in higher orders. With many width and breadth measures along with the measures of length, the factor in the highest order would probably represent body volume while stature would appear at the just lower order. This complexity obviously leads to a problem of factor identification and a requirement for exter-

nal functional relationships. In general, however, desired dispositions are not first order factors.

The test constructer can proceed with simpler statistical methods, such as item-total score correlations (internal consistency item analysis), than factor analysis but there are several problems here. An important one concerns the original item pool. If it contains substantial numbers of width and depth measures, item selection by means of the total score correlations will lead to a measure of body volume rather than of stature. Secondly, there are no criteria for deciding at what point to exclude an item from a test. If the original item pool is approximately correct, but if homogeneity standards are set too high, useful items will be excluded; if set too low items not contributing to the measurement of height, as distinguished from the correlated measures of weight, will be included.

Test constructers frequently use the steepness of age or grade curves for items administered to children as a criterion of item selection, but there are many functions that increase with increasing age. Such age curves have been commonly used in the development of tests of intelligence, but their use has again been dependent on the original item pool. Items from the pool not showing the expected relationship with age are discarded objectively, but many, many items that would have shown the expected relationships with age do not appear in the pool. For intelligence tests the choice of items for the pool has been based upon theory, tradition, and availability--and not necessarily in that order of importance.

Without recourse to an external physical measure of height the test constructer can only make use of the network of functional relationships involving his test with measures of other functions as a check on his item selection. The implications for items of the functional relationships involving total scores are quite indirect so that the change of a measurement construct by varying item

selection is a slow, arduous, and ambiguous task.

Summary of Multiple Item Function. It is hoped that the function of multiple items, as well as the difficulties inherent in their use, has become clear. The constructer of a psychological test has no physical measure to use as a criterion. He measures bits of behavior (items), each of which reflects the underlying disposition in which he is interested only to small degree. By adding the right items together he can build up the variance of the underlying disposition in the total score, but he needs to reduce random error and to spread nonrandom noise as much as possible. In the present physical analogue the test constructer is not interested in measuring finger or toe length as such. Instead his interest lies in their ability to give him a modicum of information about stature. If he includes too many measures of finger and toe length in his stature score, there will be too much variance present from factors in which he is not interested. Such bias may be present in sufficient amount to mask the information about stature. By bringing in as many indicants of stature as possible, and varying their distribution over the body as much as possible, the nonrandom noise while still present is minimized or spread over so many functions or factors other than height that the total score reflects height primarily.

For a psychological function such as verbal comprehension used in an earlier illustration the availability of a population of words from which to sample randomly, and in sufficient number, is a very important way in which to define the central function or factor that one wishes to maximize in the total score. However, item populations from which to sample are relatively rare; when items are invented, the defining of a population is at best arbitrary and at worst impossible. To distinguish between the nonrandom variance of the disposition and of noise involves a long term research operation. The most hopeful procedure if one is restricted to internal evidence is factoring the intercorrelations of carefully constructed items

in orders beyond the first. Much more often than not, first order factors represent systematic noise rather than the disposition in which the test constructor is interested, while it is only in the second or higher order that he finds the construct he is seeking to measure.

The above reasoning represents a complete break with a tradition of test construction in psychology in which high item homogeneity has been an important goal. The traditional reasoning has been that, with high homogeneity, one could infer that the test was measuring a unique, unitary function or factor. With sufficiently high homogeneity, the test becomes a Guttman scale though this is rarely attained. Scalability of a universe of items, nevertheless, became a goal toward which to strive.

This tradition leaves undefined the question as to how high the degree of homogeneity should be. This represents a formal objection to the homogeneity model, but the primary objection represented in this discussion involves the nature of test items. Items necessarily involve several kinds of components. For certain purposes one of these can be labelled the primary disposition while others are necessarily nonrandom noise and random error. Guttman scales are unobtainable except under highly restrictive, even artificial, situations.

Radical solutions should always be entertained, including the possibility that separate Guttman scales should be constructed for each source of nonerror noise, while trying to hold constant all other sources of bias. What has been noise, in other words, becomes many tests. The feasibility of this solution is in serious doubt, however, in terms of the sheer number of tests that would result. Any estimate must be labelled a guess, but it is a guess conditioned by test construction experience of psychometricians generally as well as by psychological assumptions concerning the number of possible sources, or causes, of responses to test items. The items in ability tests that are generally considered to be quite

homogeneous typically have mean levels of intercorrelations less than .20. An attitude measurement approximations to Guttman scales can be obtained with a few items having very diverse item popularities (defined statistically in a fashion similar to item difficulty) in which essentially the same question is asked with minor variations in wording. Correlations between attitude items and logical reversals of those items are not high. All in all, a guess that tens of thousands of tests would be the result is not out of line. It seems better to retain the concept of nonerror noise and to allow test constructers freedom to broaden or narrow it at will, and in accordance with scientific convenience, rather than to impose the goal of high homogeneity on all tests.

TABLE 1

Score Matrix Which Produces a Perfect

Guttman Scale

		Items									
		1	2	3	4	5	6	7	8	9	Test Score
	a	+	+	+	+	+	+	+	+	+	9
	b	+	+	+	+	+	+	+	+	0	8
	c	+	+	+	+	+	+	+	0	0	7
	d	+	+	+	+	+	+	0	0	0	6
ersons	e	+	+	+	+	+	0	0	0	0	5
	f	+	+	+	+	0	0	0	0	0	4
	g	+	+	+	0	0	0	0	0	0	3
	h	+	+	0	0	0	0	0	0	0	2
	i	+	0	0	0	0	0	0	0	0	1
	j	0	0	0	0	0	0	0	0	0	0
Difficulty Level		.9	.8	.7	.6	.5	.4	.3	.2	.1	

TABLE 2

Intercorrelations of the Items in a Perfect Guttman Scale

	1	2	3	4	5	6	7	8	9
1		.67	.31	.41	.33	.27	.22	.17	.11
2			.77	.61	.50	.41	.33	.25	.17
3				.80	.66	.54	.43	.33	.22
4					.82	.67	.54	.41	.27
5						.82	.66	.50	.33
6							.80	.61	.41
7								.77	.51
8									.67
9									

TABLE 3

Score Matrix (Schematic) in Which Measurement Error Has Been Introduced on Potentially Scalable Items

		Items									
		1	2	3	4	5	6	7	8	9	Test Score
Persons	a	+	+	+	+	+	0	+	0	+	7
	b	+	+	+	+	0	0	+	+	0	6
	c	+	+	+	0	+	+	0	0	0	5
	d	+	+	+	0	0	+	0	+	0	5
	e	+	+	0	+	+	0	+	0	0	5
	f	+	0	0	+	+	+	0	0	0	4
	g	+	0	+	+	+	0	0	0	0	4
	h	+	+	+	0	0	+	0	0	0	4
	i	+	+	0	+	0	0	0	0	0	3
	j	0	+	+	0	0	0	0	0	0	2
Difficulty Level		.9	.8	.7	.6	.5	.4	.3	.2	.1	

Abstract

A methodology has been described and illustrated for obtaining an evaluation of the importance of the factors in a particular order of factoring that does not require factoring beyond that order. For example, one can estimate the intercorrelations of the original measures with the perturbations of the first-order factors held constant or, the reverse, estimate the contribution to the intercorrelations of the original measures from the first-order factors alone. Similar operations are possible at higher orders.

EVALUATING THE IMPORTANCE OF FACTORS IN ANY GIVEN ORDER OF FACTORING

Lloyd G. Humphreys, Ledyard R. Tucker, and Peter Dachler

University of Illinois, Urbana

One of us (Humphreys, 1962) has recommended hierarchical factoring of measures of human abilities for reasons connected with a presumed gradient of importance of factors in the several orders. One indication of importance is predictive validity. Broad tests have generally higher predictive validities than narrow tests. It is also very difficult empirically to find stable differential weights for a variety of criteria for very narrow tests.

Valid objections can be raised to the evaluation of the importance of factors based upon correlations with outside criteria, but by factoring in several orders and using the Schmid-Leiman transformation (1957) to obtain a hierarchical orthogonal factor matrix an internal criterion can be obtained. The contributions to common factor variance of the several factors can be computed and compared. It occurred to us, however, that an internal criterion that did not involve higher order factoring would be useful. Such a criterion is readily available.

Mathematical Development

We shall let R stand for any matrix of intercorrelations. The subscripts 0, 1, 2, etc. will designate original intercorrelations, intercorrelations of first-order factors, intercorrelations of second-order factors, etc. Matrices of rotated factor loadings (Harman's pattern matrices on the primary axes) will be symbolized by P, also with appropriate subscripts. Estimated matrices are designated by the circumflex. Thus, we have the following well known relationships:

$$\left[\hat{R}_0 - \hat{U}_0^2\right] = P_1 R_1 P_1' \qquad (1a)$$

$$\left[\hat{R}_1 - \hat{U}_1^2\right] = P_2 R_2 P_2' \qquad (1b)$$

An evaluation of the importance of the factors in a given order is obtained by replacing the unities in the diagonal of R with the estimated communalities of the factors and multiplying as before. We symbolize the new matrix, which represents the estimated intercorrelations at a lower order with the perturbations of the factors at the next higher order removed, in a fashion analogous to partial correlations. Thus we have:

$$\hat{R}_{0.1} = P_1 (R_1 - \hat{U}_1^2) P_1' \qquad (2a)$$

$$\hat{R}_{1.2} = P_2 (R_2 - \hat{U}_2^2) P_2' \qquad (2b)$$

Other matrices of interest can be derived immediately from the above. The direct contribution of the first-order factors, in contrast to the control of their effects, is given by the following:

$$[\hat{R}_0 - \hat{U}_0^2] - \hat{R}_{0.1} = P_1 \hat{U}_1^2 P_1' \qquad (3a)$$

$$[\hat{R}_1 - \hat{U}_1^2] - \hat{R}_{1.2} = P_2 \hat{U}_2^2 P_2' \qquad (3b)$$

These direct contributions of the factors can be designated as $\hat{R}_{0.2,3..k}$ and $\hat{R}_{1.3..k}$ also in a fashion analogous to partial correlations to indicate that the effects of higher order factors have been removed. Thus the entries in the matrix $\hat{R}_{0.2,3..k}$ indicate the contributions of the first order factors only to the correlations among the original measures. It also follows that $\hat{R}_{0.1,2,3..k} = \hat{U}_0^2$, which represents another example of the symbolism.

It has been suggested that low $\hat{U}^2$ values for factors may be obtained when there is substantial capitalization upon chance (Horn, 1966), or when rotations have been contrived to force data into a particular structure (Humphreys, 1967). Low values of $\hat{U}^2$ can also be obtained legitimately and objectively from the nature of the data. Whatever the basis may be, if the matrices of contribution to correlations obtained in the fashion of Formula 3a contain values close to zero, the first order factors in question can be considered relatively unimportant. A "real"

factor may make only a minor contribution to covariation. A similar statement may be made for $\hat{R}_{1.3..k}$ and second order factors.

It is possible to estimate communalities for factors in the several ways that one estimates communalities for the original measures, but under certain conditions squared multiples have much to recommend them. Multiple correlations, depending as they do on the entire matrix of intercorrelations, are more stable than many other communality estimates. Although squared multiples are lower bound estimates only in the population of observations and approach "true" communalities only as the population of measures is approached, they tend not to be seriously in error when the number of measures, or factors as in the present case, is moderately large and when the number of observations is much larger.

When the use of squared multiples is appropriate, it is unnecessary to factor in a higher order in order to use formulas 2 and 3 in evaluating factor importance. Thus the error of estimate variance, symbolized as S^2, can be substituted for $\hat{U}^2$ in formulas 2 and 3.

$$\hat{R}_{0.1} = P_1 (R_1 - S^2) P_1' \qquad (4)$$

$$\hat{R}_{1.2,3..k} = P_1 S^2 P_1 \qquad (5)$$

We can now, in turn, let V = PS, which leads to the following relationship:

$$\hat{R}_{0.2,3..k} = V_1 V_1' \qquad (6)$$

The matrix V contains the projection of the measures on the normals to the hyperplanes (Harman's reference vector structure). Consequently, these projections can also be used to evaluate the contributions of the factors. Small correlations with the reference vectors of measures of high communality indicate that the important factors are in a higher order.

Illustrations of the Procedures

In order to illustrate these procedures we turned to published data. The

Adkins and Lyerly analysis of reasoning tests (1952) contains sufficient numbers of factors in the first order to allow use of squared multiples as communality estimates. The same is true of the first-order factors in Cattell's study of fluid and crystallized ability (1963).

It is not feasible to present large estimated intercorrelational matrices. Other indices must be found. Distributions of the diagonals of the several matrices constitute one compact way of describing the contributions of first order and higher order factors. Intercorrelations of selected variables can also be shown as more concrete illustrations of the effects of first and higher order factors.

Table 1 contains the distributions described above along with means and standard deviations. The first order factors in the Adkins and Lyerly data are responsible for a higher percentage of the variance than the first order factors in the Cattell data. It should also be borne in mind in interpreting these results that the available communality would ordinarily be spread over many more first-order factors than all higher order factors combined. Furthermore, several of the Cattell first-order factors are specifics so that diagonals on $R_{0.2,3...k}$ are, in a sense, inflated.

The lesser importance of Cattell's first-order factors, it should also be noted, is not critical with respect to his conclusions. His study was designed for higher order factoring. In other studies, however, in which the first-order factors are of prime importance to the investigator, the technique here being illustrated is a desirable, even necessary, check on the conclusions reached.

In both of the two sets of data the correlation between the two estimates of the diagonal is only moderately negative. This means that the effects of controlling first-order factors or higher order factors are far from homogeneous with

respect to the measures. Subsets of measures are differentially affected by factors in the several orders. Table 2, for example, contains estimated and obtained intercorrelations of selected verbal tests from Adkins and Lyerly. Obtained correlations are below the diagonal while the estimated contributions to the correlations of first-order factors alone and higher order factors alone are above the diagonal. Within each cell above the diagonal the upper value represents $R_{0.2,3..k}$, the lower value $R_{0.1}$. Thus we see that the intercorrelations of the verbal tests in this analysis tend to be explained more by higher order factors than by first-order factors.

Table 3 contains similar data for the Primary Mental Ability measures used by Cattell. First-order factors contribute only to the correlations between the parallel forms while the higher order factors account, as one would expect, for the intercorrelations of the different tests. The Fluency test for which no parallel form was available defined a specific factor in the Cattell analysis. This is clearly seen in the correlations presented. Furthermore, 39 of the possible 40 correlations with other variables in the full $R_{0.2,3..k}$ matrix are smaller than .10 for the Fluency measure, and the fortieth is less than .20.

Summary and Conclusions

A methodology has been described and illustrated for obtaining an evaluation of the importance of the factors in a particular order of factoring that does not require factoring beyond that order. For example, one can estimate the intercorrelations of the original measures with the perturbations of the first-order factors held constant or, the reverse, estimate the contribution to the intercorrelations of the original measures from the first-order factors alone. Similar operations are possible at higher orders.

An estimate of communality of the factors at a given level is required in order to estimate correlations at the lower level. When many factors are involved, squared multiples can be used for this purpose. Under these circumstances, also, the importance of the factors can be gauged by the size of the correlations of the original measures, or factors, with the reference vector structure for those measures or factors.

References

Adkins, D. C. and Lyerly, S. B. Factor analysis of reasoning tests. 1952, Chapel Hill: The University of North Carolina Press, 122 pp.

Cattell, R. B. Theory of fluid and crystallized intelligence: a critical experiment. Journal of Educational Psychology, 1963, 54, 1-22.

Horn, J. L. On subjectivity in factor analysis. Educational and Psychological Measurement, 1967, 27, 811-820.

Humphreys, L. G. The organization of human abilities. American Psychologist, 1962, 17, 475-483.

Humphreys, L. G. Critique of Cattell's "Theory of fluid and crystalized: a critical experiment." Journal of Educational Psychology, 1967, 58, 129-136.

Schmid, J. and Leiman, J. The development of hierarchical factor solutions. Psychometrika, 1957, 22, 53-61.

Footnotes

1. This research was supported by the Office of Naval Research under contracts N 00014-67-A-0305-0003, Ledyard R. Tucker, principal investigator, and N 00014-67-A-0305-0012, Lloyd G. Humphreys, principal investigator.

Table 1

Distributions of Diagonal Values in $R_{0.1}$ and $R_{0.2,3...k}$ from Two Separate Analyses

	Adkins and Lyerly (1952)		Cattell (1963)	
	$R_{0.1}$	$R_{0.2,3...k}$	$R_{0.1}$	$R_{0.2,3...k}$
65		1		
60		1		1
55		1		0
50	1	2		1
45	1	6		5
40	3	13	2	4
35	7	14	4	5
30	9	16	8	6
25	7	5	15	9
20	13	5	7	7
15	9	2	2	3
10	5		3	
05	3			
00	5			
-05	2			
-10	0			
-15	1			
$\bar{X}$	.22	.37	.28	.33
S	.13	.10	.10	.11

Table 2

Comparison of Obtained and Estimated Intercorrelations

of Selected Verbal Tests from Adkins and Lyerly (1952)*

	49	50	59	60	61	62
49 Reading 1		30 32	26 32	21 38	21 36	44 29
50 Reading 2	65		21 36	21 38	23 37	27 33
59 Verbal Analogies	57	54		17 41	20 41	28 34
60 Verbal Classification 1	58	59	58		30 48	21 41
61 Verbal Classification 2	56	60	61	87		21 40
62 Vocabulary	74	65	65	63	62	

* Measures are numbered as in Adkins and Lyerly. Entries below the diagonal are observed intercorrelations; the upper one of the pair of entries above the diagonal is from $R_{0.2,3..k}$ (first order factor contributions), and the lower one is from $R_{0.1}$ (higher order factor contributions).

Table 3

Intercorrelations of PMA Measures

from Cattell (1963)*

	1	2	3	4	5	6	7	8	9
1 Verbal 1		48 38	04 27	05 26	00 43	02 39	-02 36	-02 34	08 36
2 Verbal 2	86		00 29	00 28	00 43	02 40	00 39	-02 36	09 37
3 Space 1	30	30		53 26	-02 24	00 24	-04 19	01 13	05 20
4 Space 2	32	27	79		-03 25	03 23	-04 20	01 14	05 16
5 Reasoning 1	41	42	21	23		40 38	02 41	04 41	03 34
6 Reasoning 2	42	41	25	25	77		-02 39	01 38	05 30
7 Numerical 1	34	37	23	16	40	37		44 33	04 39
8 Numerical 2	32	33	19	14	43	39	78		04 36
9 Fluency	44	45	17	22	36	33	42	40	

* Entries below the diagonal are observed intercorrelations; the upper one of the pair of entries above the diagonal is from $R_{0.2,3..k}$ (first order factor contributions), and the lower one is from $R_{0.1}$ (higher order factor contributions).

Abstract

A major conclusion of the 1947 Scottish survey of intelligence was that there had been a gain since 1932 on the group test but no gain on the individual Stanford-Binet test. This conclusion is marred, however, by the use of regression methods of equating the 1916 and 1937 editions of the individual test for which only 89 cases were available. Avoidance of the sample of 89 cases who had been administered both editions of the individual test by the use of the equipercentile method of equation reveals parallel gains for both the group and the individual test. There is no need to qualify the conclusion that a small increase in intelligence among Scottish school children occurred between 1933 and 1947.

Footnote to the Scottish Survey of Intelligence

Lloyd G. Humphreys

University of Illinois

The results of the 1947 Scottish survey of intelligence (Scottish Council for Research in Education, 1949) were somewhat ambiguous, as reported, with respect to a gain in intelligence between 1932 and 1947. The group test, which was the principal survey instrument, showed a gain. The 1916 and the 1937 editions of the Stanford-Binet were also administered in 1933 and 1947, respectively, to 500 or more students of each sex. When the two editions were equated, by a procedure which will be described briefly below, the authors reported a slight loss for girls. Overall the mean intelligence quotients were almost precisely the same. The conflict in results between the group test and the individual tests has sometimes been interpreted as meaning that "real" intelligence did not change.

The procedures used in equating the two versions of the individual test are, however, open to question. First, the authors used a regression method for equating group and individual test scores separately for the two sexes. Then they used a regression method for equating the scores on the separate individual tests. The equation of group to individual test involved 500 or more of the special cases for each sex, but the final equation of the two individual tests involved only 89 cases of both sexes combined.

A good case can be made against the regression method of equating scores on two tests each of which is supposedly measuring basically the same function. An even better case can be made for forgetting about the 89 cases who had been given both versions of the individual test and base the equation of the two on their relationship to the group test. An N of 1000 or more for this step in the procedure has clear advantages over an N of 89 but use of the large N requires the equipercentile method of conversion.

There are advantages to the use of the equipercentile method over and beyond its applicability to the data based upon the large N. It requires only ordinal scales of measurement and is thus independent of the shape of the regression. It is also independent of attentuation in the slope of the regression introduced by measurement error. It does require the assumption that the two measures are equally valid measures of the trait, but the regression method requires a different and at least equally difficult assumption that one of the two measures can be considered the criterion measure of the trait.

Table 1 contains comparable scores for boys and girls for each level of the group test on the 1916 and 1937 editions of the Stanford-Binet. Also included are the differences between the intelligence quotients for each level of the group test. When these differences are weighted by their respective Ns and averaged, it is seen that the mean difference in intelligence quotients for boys between the 1916 and 1937 editions is 3.14 units. The comparable figure for girls is -.99. There is clearly an interaction between sex and the two editions of the Stanford-Binet when the conversion is based upon the group test common to both testing periods. This same interaction is also seen in the results from the 6-day sample in which girls are slightly superior to boys on the group test and significantly inferior to boys on the 1937 revision. For some reason, Scottish girls seemed to be at a relative disadvantage on the revised Stanford-Binet in spite of the near equality of the sexes in the standardization samples. The latter were, of course, drawn entirely from the United States.

There are two ways in which the equipercentile method can be applied. Equipercentile conversions can be computed between the group test and each individual test, and then the individual tests can be converted to each other. Alternatively, the published regression conversions of individual test on group test can be

accepted, but an equipercentile conversion can be substituted for the published regression conversion between the two individual tests. The latter will be designated the "mixed" method.

Table 2 presents the published data on group test and the regression conversions of the individual tests in the first three numbered rows. These data are followed by equipercentile conversions computed by the present writer. The means in lines 2 and 4 which represent different estimates of the population I. Q.s on the individual tests are generally comparable though the equipercentile values are somewhat lower. This difference is a regression phenomenon, arising from the lack of perfect correlation between group and individual test. Conversions of 1916 I. Q.s into 1937 I. Q.s are contained in lines 3, 5, and 7. The two variations of the equipercentile method are in general agreement in showing a gain in I. Q. for girls, but both depart radically from the regression results. Furthermore, when the 1932 results are presented in the units of the 1937 revision of the Stanford-Binet (lines 6 and 8), the discrepancy between the results for the two sexes is very marked.

The weak conclusion that can be drawn from this analysis is that the previous outcome of no gain on an individual test of intelligence between 1932 and 1947 is questionable since a different and supportable methodology demonstrates a gain. The strong conclusion, which accepts the superiority of the equipercentile methodology for problems of this type, is that gains on group and individual tests are approximately parallel and that the gain was greater for girls than for boys. From the latter point of view there is no need to qualify the conclusion that a small increase in intelligence among Scottish school children occurred between 1933 and 1947. One is strongly tempted, furthermore, to predict that the gain would have been larger without the disrupting effects on education of World War II.

Even so the results are well in line with Tuddenham's demonstration of an increase in intelligence among men in the United States between World War I and World War II (1948).

Table 1

Equivalent Scores on the 1916 and 1937 Editions of the Stanford-Binet for Various Levels of the Group Test

	Boys			Girls		
Group test	1937 ed.	1916 ed.	Difference	1937 ed.	1916 ed.	Difference
69.5	167.83	149.17	18.66	154.50	149.50	5.00
64.5	151.17	140.12	11.05	149.50	139.50	10.00
59.5	134.50	128.46	6.04	133.71	132.50	1.21
54.5	125.02	121.53	3.69	122.68	119.71	2.97
49.5	116.06	112.94	3.12	113.72	112.36	1.38
44.5	108.46	106.37	2.09	107.10	107.05	.05
39.5	103.96	100.48	3.48	99.27	100.01	-.74
34.5	100.11	95.33	4.78	94.01	94.21	-.20
29.5	95.77	90.88	4.89	88.43	90.22	-1.79
24.5	89.85	87.38	2.47	82.68	86.17	-3.49
19.5	85.96	85.26	.70	79.27	82.28	-3.01
14.5	82.09	81.57	.52	75.52	79.05	-3.53
9.5	77.71	77.71	.00	72.17	75.41	-3.24
4.5	73.79	71.64	2.15	67.23	70.05	-2.82
-.5	54.50	54.50	.00	49.50	64.50	-15.00
Weighted Mean Difference			3.14			-.99

Table 2

Summary of Results for 1932 and 1947

		Boys		Girls	
		1932	1947	1932	1947
(1)	Group Test	34.503	35.880	34.409	37.622
	Regression Conversions				
(2)	Individual on Group	99.86	103.68	98.56	100.75
(3)	1916 on 1937		100.48		97.89
	Equipercentile Conversions				
(4)	Group to Individual	98.29	103.52	97.67	100.61
(5)	1937 to 1916		99.91		101.49
(6)	1916 to 1937	101.70		96.33	
	Mixed Conversions*				
(7)	1937 to 1916		100.54		101.74
(8)	1916 to 1937	103.00		97.57	

*Utilizes the regression conversion of each individual test on the group test (line 2) as the first step, but uses an equipercentile conversion as the second step.

References

Scottish Council for Research in Education, The Trend of Scottish Intelligence, 1949, University of London Press, London, 151 pp.

Tuddenham, R. D. Soldier intelligence in World Wars I and II, American Psychologist, 1948, 3, 54-56.

May 1970

OFFICE OF NAVAL RESEARCH

PERSONNEL AND TRAINING RESEARCH PROGRAMS (CODE 458)

DISTRIBUTION LIST

CONTRACT NO. Noo14-67-A-0305-0012

CONTRACTOR University of Illinois
Lloyd G. Humphreys

NAVY

4 Chief of Naval Research
Code 458
Department of the Navy
Arlington, Virginia 22217

1 Director
ONR Branch Office
495 Summer Street
Boston, Massachusetts 02210

1 Director
ONR Branch Office
219 South Dearborn Street
Chicago, Illinois 60604

1 Director
ONR Branch Office
1030 East Green Street
Pasadena, California 91101

6 Director, Naval Research Laboratory
Washington, D.C. 20390
ATTN: Library, Code 2029 (ONRL)

1 Office of Naval Research
Area Office
207 West Summer Street
New York, New York 10011

1 Office of Naval Research
Area Office
1076 Mission Street
San Francisco, California 94103

6 Director
Naval Research Laboratory
Washington, D.C. 20390
ATTN: Technical Information Division

20 Defense Documentation Center
Cameron Station, Building 5
5010 Duke Street
Alexandria, Virginia 22314

1 Commanding Officer
Service School Command
U.S. Naval Training Center
San Diego, California 92133

3 Commanding Officer
Naval Personnel and Training
Research Laboratory
San Diego, California 92152

1 Commanding Officer
Naval Medical Neuropsychiatric
Research Unit
San Diego, California 92152

1 Commanding Officer
Naval Air Technical
Training Center
Jacksonville, Florida 32213

1 Dr. James J. Regan, Code 55
Naval Training Device Center
Orlando, Florida 32813

1 Technical Library
U.S. Naval Weapons Laboratory
Dahlgren, Virginia 22448

1 Research Director, Code 06
Research and Evaluation
Department
U.S. Naval Examining Center
Building 2711 - Green Bay Area
Great Lakes, Illinois 60088
ATTN: C.S. Winiewicz

1 Chairman
Behavioral Science Department
Naval Command and
Management Division
U.S. Naval Academy
Luce Hall
Annapolis, Maryland 21402

1 Chairman
Management Science Department
Naval Command and Management Div.
U.S. Naval Academy
Luce Hall
Annapolis, Maryland 21402

1 Dr. A. L. Slafkosky
Scientific Advisor (Code AX)
Commandant of the Marine Corps
Washington, D.C. 20380

1 Behavioral Sciences Department
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Maryland 20014

1 Commanding Officer
Naval Medical Field
Research Laboratory
Camp Lejeune, North Carolina 28542

1 Director, Aerospace Crew
Equipment Department
Naval Air Development Center
Johnsville
Warminster, Pennsylvania 18974

1 Chief, Naval Air
Technical Training
Naval Air Station
Memphis, Tennessee 38115

1 Director, Education and
Training Sciences Department
Naval Medical Research Institute
National Naval Medical Center
Building 142
Bethesda, Maryland 20014

1 Commander, Submarine Development
Group TWO
Fleet Post Office
New York, New York 09501

1 Commander, Operational
Test and Evaluation Force
U.S. Naval Base
Norfolk, Virginia 23511

1 Office of Civilian Manpower
Management, Technical
Training Branch (Code 024)
Department of the Navy
Washington, D.C. 20390

1 Chief of Naval Operations
(Op-07TL)
Department of the Navy
Washington, D.C. 20350

1 Chief of Naval Material (MAT 031M)
Room 1323, Main Navy Building
Washington, D.C. 20360

1 Mr. George N. Graine
Naval Ship Systems Command
(SHIPS 03H)
Department of the Navy
Washington D.C. 20360

1 Chief, Bureau of
Medicine and Surgery
Code 513
Washington, D.C. 20390

1 Chief, Bureau of Medicine
and Surgery
Research Division (Code 713)
Department of the Navy
Washington, D.C. 20390

9 Technical Library (Pers-11b)
Bureau of Naval Personnel
Department of the Navy
Washington, D.C. 20370

3 Personnel Research and
Development Laboratory
Washington Navy Yard, Building 200
Washington, D.C. 20390
ATTN: Library, Room 3307

1 Commander, Naval Air Systems
Command, Navy Department
AIR-4132
Washington, D.C. 20360

1 Commandant of the Marine Corps
Headquarters, U.S. Marine Corps
Code A01B
Washington, D.C. 20380

1 Technical Library
Naval Ship Systems Command
Main Navy Building, Room 1532
Washington, D.C. 20360

1 Mr. Philip Rochlin, Head
Technical Library Branch
Naval Ordnance Station
Indian Head, Maryland 20640

1 Library, Code 0212
Naval Postgraduate School
Monterey, California 93940

1 Technical Reference Library
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Maryland 20014

1 Scientific Advisory Team (Code 71)
Staff, COMASWFORLANT
Norfolk, Virginia 23511

1 Education & Training
Developments Staff
Personnel Research &
Development Laboratory
Washington Navy Yard, Bldg. 200
Washington, D.C. 20390

1 Dr. Don H. Coombs, Co-Director
ERIC Clearinghouse
Stanford University
Palo Alto, California 94305

1 ERIC Clearinghouse on
Educational Media and Technology
Stanford University
Stanford, California 94305

1 ERIC Clearinghouse on Vocational
and Technical Education
Ohio State University
1900 Kenny Road
Columbus, Ohio 43210
ATTN: Acquisition Specialist

1 Lt. Col. F. R. Ratliff
Office of the Assistant
Secretary of Defense (M&RU)
The Pentagon, Room 3D960
Washington, D.C. 20301

1 Dr. Ralph R. Canter
Military Manpower
Research Coordinator
OASD (M&RA) MR&U
The Pentagon, Room 3D960
Washington, D.C. 20301

1 Deputy Director
Office of Civilian Manpower
Management
Department of the Navy
Washington, D.C. 20390

1 Chief, Naval Air
Reserve Training
Naval Air Station, Box 1
Glenview, Illinois 60026

1 Technical Library
Naval Training Device Center
Orlando, Florida 32813

1 U.S. Naval Submarine
Medical Center
P. O. Box 600, Naval Submarine Base
Groton, Connecticut 06340
ATTN: Dr. B. B. Weybew

1 Dr. Robert F. Lockman
Center for Naval Analyses
1401 Wilson Boulevard
Arlington, Virginia 22209